Bias reduction in the estimation of mutual information

Jie Zhu, Jean-Jacques Bellanger, Huazhong Shu, Chunfeng Yang, Régine Le Bouquin Jeannès

Bias reduction in the estimation of mutual information

Jie Zhu,[1,2,3] Jean-Jacques Bellanger,[1,2] Huazhong Shu,[3,4] Chunfeng Yang,[3,4] and Régine Le Bouquin Jeannès[1,2,3,*]

[1] *INSERM, U 1099, Rennes, F-35000, France*
[2] *Université de Rennes 1, LTSI, F-35000, France*
[3] *Centre de Recherche en Information Biomédicale sino-français (CRIBs), Rennes, France*
[4] *LIST, School of Computer Science and Engineering, Southeast University, Nanjing, China*

(Dated: September 25, 2014)

This paper deals with the control of bias estimation when estimating mutual information from nonparametric approach. We focus on continuously distributed random data and the estimators we developed are based on nonparametric k-nearest neighbor approach for arbitrary metrics. Using a multidimensional Taylor series expansion, a general relationship between the estimation error bias and neighboring size for plug-in entropy estimator is established without any assumption on the data for two different norms. The theoretical analysis based on the maximum norm developed coincides with the experimental results drawn from numerical tests made by Kraskov *et al.*, Phys. Rev. E **69**. 066138 (2004). To further validate the novel relation, a weighted linear combination of distinct mutual information estimators is proposed and, using simulated signals, the comparison of different strategies allows for corroborating the theoretical analysis.

PACS numbers: 89.70.Cf,87.19.lo,02.50.-r

I. INTRODUCTION

Mutual Information (MI) is a widely used information theoretical independence measurement which has received particular attention during the past few decades. However, the estimation of MI remains a tough task while carried out on finite sample length signals, for example in the field of neuroscience, where getting large amounts of stationary data is problematical. More precisely, let (X,Y) be a pair of multidimensional random variables with a continuous distribution specified by a joint probability density $p_{X,Y}$ with marginal densities p_X and p_Y. The joint and marginal entropies, namely $\mathcal{H}(X,Y)$, $\mathcal{H}(X)$ and $\mathcal{H}(Y)$, respectively linked to (X,Y), X and Y, are defined as $\mathcal{H}(X,Y) = -\mathrm{E}\left[\log p_{X,Y}\left(X,Y\right)\right]$, $\mathcal{H}(X) = -\mathrm{E}\left[\log p_X\left(X\right)\right]$ and $\mathcal{H}(Y) = -\mathrm{E}\left[\log p_Y\left(Y\right)\right]$. Mutual information between X and Y is then defined as [1]

$$\begin{aligned}\mathcal{I}(X,Y) &= \int \log\left[\frac{p_{X,Y}\left(x,y\right)}{p_X\left(x\right)p_Y\left(y\right)}\right] p_{X,Y}\left(x,y\right)\mathrm{d}x\mathrm{d}y \\ &= \mathcal{H}(X)+\mathcal{H}(Y)-\mathcal{H}(X,Y).\end{aligned} \tag{1}$$

According to Eq. (1), MI estimation could be simply obtained by estimating three individual entropies separately and then summing them. In this way, it is possible to choose relation-specific parameters to cancel out the bias errors in individual estimations to avoid an adverse accumulation of errors. To this end, Kraskov *et al.* [2] proposed to use a common neighboring size for both joint and marginal spaces when selecting nearest neighbors. This strategy consisted in fixing the number of neighbors in the joint space $\mathcal{S}_Z$ $[Z=(X,Y)]$, then projecting the resulting distance into the marginal spaces $\mathcal{S}_X$ and $\mathcal{S}_Y$. Following this idea, two different MI estimators giving comparable results were proposed [2]:

$$\widehat{\mathcal{I}\left(X,Y\right)}_{\mathrm{K1}} = \psi\left(k\right) - \left\langle\psi\left(n_X+1\right)+\psi\left(n_Y+1\right)\right\rangle + \psi\left(n\right) \tag{2}$$

and

$$\widehat{\mathcal{I}\left(X,Y\right)}_{\mathrm{K2}} = \psi\left(k\right) - \frac{1}{k} - \left\langle\psi\left(n_X\right)+\psi\left(n_Y\right)\right\rangle + \psi\left(n\right), \tag{3}$$

where n is the signal length, k the number of neighbors, $\psi\left(\cdot\right)$ denotes the digamma function, the symbol $\langle\cdot\rangle$ stands for an averaging on a sample data set, n_X and n_Y are the numbers of points which fall into the resulting distances in the marginal spaces $\mathcal{S}_X$ and $\mathcal{S}_Y$ respectively.

In [2], the effectiveness of this strategy to reduce bias is attested through numerical experiments. This strategy has also been extended to the calculation of other information theory functionals, such as divergence [3] or conditional mutual information [4]. In [2], the following interesting conjecture has been raised from simulation results:

$$\mathrm{E}\left[\widehat{\mathcal{I}\left(X,Y\right)}_{\mathrm{K1}}\right] = \mathrm{E}\left[\widehat{\mathcal{I}\left(X,Y\right)}_{\mathrm{K2}}\right] = 0, \text{ iif } \mathcal{I}\left(X,Y\right) = 0, \tag{4}$$

where the expectation is computed from the joint probability distribution of the data sample including all the observed occurrences of (X,Y).

In the present work, we propose to give some theoretical explanations to justify this result before developing a new estimator.

*Electronic address: regine.le-bouquin-jeannes@univ-rennes1.fr

$$\frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)} \approx p_X(x) + \left[\frac{\partial p_X(x)}{\partial x}\right]^T \frac{1}{v(x)} \int_{\mathcal{L}(x)} (y-x)\mathrm{d}y + \frac{1}{2v(x)} \int_{\mathcal{L}(x)} (y-x)^T \left[\frac{\partial^2 p_X(x)}{\partial x^2}\right] (y-x)\mathrm{d}y \tag{12}$$

$$\begin{aligned} \frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)} &\approx p_X(x) + \frac{1}{2v(x)} \int_{\mathcal{L}(x)} (y-x)^T \left[\frac{\partial^2 p_X(x)}{\partial x^2}\right] (y-x)\mathrm{d}y \\ &= p_X(x) + \frac{1}{2v(x)} \mathrm{tr}\left\{ \left[\int_{\mathcal{L}(x)} (y-x)(y-x)^T \mathrm{d}y \frac{\partial^2 p_X(x)}{\partial x^2} \right] \right\} \end{aligned} \tag{13}$$

II. METHODS AND MATERIALS

A. New bias expression for the plug-in entropy estimator

Let us consider a d_X dimensional random variable X whose outcomes are in $\mathbb{R}^{d_X}$. If for any x in $\mathbb{R}^{d_X}$, $\mathcal{L}(x)$ stands for a small region around x, we introduce the volume (Lebesgue measure) $v(x) = \int_{\mathcal{L}(x)} \mathrm{d}z$ of $\mathcal{L}(x)$ and the probability density function $p_X(x)$ to specify the probability measure P_X on this space. In most existing density estimation algorithms, including either KDE (Kernel Density Estimation) or kNN (k-Nearest Neighbor), $p_X(x)$ is estimated as

$$\widehat{p_X}(x) = \frac{\widehat{P\left[X \in \mathcal{L}(x)\right]}}{v(x)} = \frac{\widehat{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}}{v(x)}, \tag{5}$$

where $\widehat{P\left[X \in \mathcal{L}(x)\right]}$ corresponds to an estimation of the probability that X belongs to the volume $v(x)$. If we assume that $P\left[X \in \mathcal{L}(x)\right]$ is perfectly known [but not $p_X(x)$], we can use the following approximation

$$\begin{aligned} \log p_X(x) &\approx \log\left\{\frac{P\left[X \in \mathcal{L}\left(x\right)\right]}{v\left(x\right)}\right\} \\ &= \log\left[\frac{\int_{\mathcal{L}(x)} p_X\left(y\right)\mathrm{d}y}{v\left(x\right)}\right]. \end{aligned} \tag{6}$$

Given Eq. (5), an estimation $\widehat{\log p_X(x)}$ of $\log p_X(x)$ is introduced

$$\begin{aligned} \widehat{\log p_X(x)} &= \log \widehat{p_X}\left(x\right) \\ &= \log \frac{\widehat{P\left[X \in \mathcal{L}(x)\right]}}{v(x)} \\ &= \log\left[\frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)} + \varepsilon\right], \end{aligned} \tag{7}$$

where the random estimation error ε given by

$$\varepsilon = \frac{\widehat{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}}{v(x)} - \frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)} \tag{8}$$

is zero mean when $\widehat{P\left[X \in \mathcal{L}(x)\right]}$ is unbiased.

From observations X_i (random variables) issued from P_X, the corresponding differential entropy $\mathcal{H}(X)$ can be estimated as

$$\widehat{\mathcal{H}(X)} = -\frac{1}{n}\sum_{i=1}^{n} \widehat{\log p_X}(X_i), \tag{9}$$

where n is the number of data used in the averaging. Then, we approximate the probability density $p_X(y)$ using a second-order Taylor approximation around x,

$$\begin{aligned} p_X(y) \approx\ & p_X(x) + \left[\frac{\partial p_X(x)}{\partial x}\right]^T (y-x) \\ & + \frac{1}{2}(y-x)^T \left[\frac{\partial^2 p_X(x)}{\partial x^2}\right] (y-x), \end{aligned} \tag{10}$$

with the superscript T standing for matrix transposition, and analyze the bias of $\widehat{\mathcal{H}(X)}$ with

$$\begin{aligned} \widehat{\mathcal{H}(X)} &= -\frac{1}{n}\sum_{i=1}^{n} \log \widehat{p_X}(X_i) \\ &= -\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{\int_{\mathcal{L}(X_i)} p_X(y)\mathrm{d}y}{v(X_i)} + \varepsilon_i\right], \end{aligned} \tag{11}$$

where the index i refers to the sample number. Integrating Eq. (10) on both sides and dividing by $v(x)$, we get Eq. (12).

If $\mathcal{L}(x)$ admits x as a center of symmetry, then $\int_{\mathcal{L}(x)}(y-x)\mathrm{d}y = 0$ and the first order term on the right hand side of Eq. (12) is zero. According to matrix properties [5], Eq. (12) can be transformed into Eq. (13), where $\mathrm{tr}\left(\cdot\right)$ stands for the trace operator [note that $\int_{\mathcal{L}(x)} \left(y-x\right)\left(y-x\right)^T \mathrm{d}y$ is a diagonal matrix].

Finally, the estimator $\widehat{\log p_X}(x)$ of $\log p_X(x)$ can be approximated by Eq. (14), where the term $\left[\frac{1}{p_X(x)} \cdot \varepsilon\right]$ is zero mean.

The bias $\mathcal{B}_X$ in $\widehat{\mathcal{H}(X)}$ is approximated by the second term in the right hand side of Eq. (14) and used as a correcting term. To build $\mathcal{L}(x)$ which admits x as a center of symmetry, we retain two norms, the Euclidean norm ($\|\cdot\| = \|\cdot\|_{\mathrm{E}}$) and the maximum norm ($\|\cdot\| = \|\cdot\|_{\mathrm{M}}$) such

$$\log\left[\frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)}+\varepsilon\right] \approx \log\left(p_X(x)+\frac{1}{2v(x)}\mathrm{tr}\left\{\left[\int_{\mathcal{L}(x)}(y-x)(y-x)^T\mathrm{d}y\right]\left[\frac{\partial^2 p_X(x)}{\partial x^2}\right]\right\}+\varepsilon\right)$$
$$\approx \log p_X(x) + \underbrace{\frac{1}{p_X(x)}\frac{1}{2v(x)}\mathrm{tr}\left\{\left[\int_{\mathcal{L}(x)}(y-x)(y-x)^T\mathrm{d}y\right]\left[\frac{\partial^2 p_X(x)}{\partial x^2}\right]\right\}}_{\approx \mathcal{B}_X} + \frac{1}{p_X(x)}\varepsilon \tag{14}$$

$$\widehat{\mathcal{I}(X,Y)} = -\frac{1}{n}\sum_{i=1}^{n}\left\{\log\widehat{p_X}(x_i)+\log\widehat{p_Y}(y_i)-\log\widehat{p_Z}(z_i)-[\mathcal{B}_X(x_i)+\mathcal{B}_Y(y_i)-\mathcal{B}_Z(z_i)]\right\} \tag{17}$$

$$\frac{1}{p_Z(z_i)}\cdot\mathrm{tr}\left[\frac{\partial^2 p_Z(z_i)}{\partial z^2}\right] = \frac{1}{p_X(x_i)}\cdot\mathrm{tr}\left[\frac{\partial^2 p_X(x_i)}{\partial x^2}\right]+\frac{1}{p_Y(y_i)}\cdot\mathrm{tr}\left[\frac{\partial^2 p_Y(y_i)}{\partial y^2}\right] \tag{18}$$

$$\widehat{\mathcal{I}(X,Y)}^{k}_{\text{basic}} = \widehat{\mathcal{H}(X)}_{\text{basic}}+\widehat{\mathcal{H}(Y)}_{\text{basic}}-\widehat{\mathcal{H}(Z)}_{\text{basic}} = -\frac{1}{n}\sum_{i=1}^{n}\left[\log\frac{k\,(x_i)}{n\cdot v(x_i)}+\log\frac{k\,(y_i)}{n\cdot v(y_i)}-\log\frac{k(z_i)}{n\cdot v(z_i)}\right] \tag{23}$$

---

that $\mathcal{L}(x)=\{y:\|y-x\|\leq\mathcal{R}(x)\}$ corresponding respectively to a standard ball and to a d_X dimensional cube. Consequently, the value $\mathcal{R}(x)$ fixes respectively the radius of the ball or the half of the edge length of the cube.

After calculation, using the Euclidean norm [5], we get

$$\mathcal{B}_X(x)\approx\frac{\mathcal{R}^2(x)}{2(d_X+2)}\cdot\frac{1}{p_X(x)}\cdot\mathrm{tr}\left[\frac{\partial^2 p_X(x)}{\partial x^2}\right]. \tag{15}$$

Similarly, using the maximum norm distance, we get

$$\mathcal{B}_X(x)\approx\frac{\mathcal{R}^2(x)}{6}\cdot\frac{1}{p_X(x)}\cdot\mathrm{tr}\left[\frac{\partial^2 p_X(x)}{\partial x^2}\right]. \tag{16}$$

Note that, with the second order approximation, the bias $\mathcal{B}_X$ increases with larger $\mathcal{R}(x)$ whatever the norm.

B. Bias reduction of MI estimator based on the new bias expression

If we come back to the estimation of mutual information, with the help of Eq. (14), by subtracting the bias terms, we propose the estimation given by Eq. (17).

Consider the ith data point, if the signals X and Y are independent, i.e., $p_Z(z)=p_X(x)p_Y(y)$, with $Z=(X,Y)$, we obtain Eq. (18).

In this case, we impose relationship-specific distances for different entropy estimations in Eq. (1) to cancel out the bias, i.e.,

$$\mathcal{B}_X(x_i)+\mathcal{B}_Y(y_i)-\mathcal{B}_Z(z_i)=0. \tag{19}$$

With the Euclidean norm, it yields to

$$\mathcal{R}(x_i)=\sqrt{\frac{d_X+2}{d_Z+2}}\cdot\mathcal{R}(z_i)\ \text{ and }\ \mathcal{R}(y_i)=\sqrt{\frac{d_Y+2}{d_Z+2}}\cdot\mathcal{R}(z_i), \tag{20}$$

where $\mathcal{R}(x_i)$, $\mathcal{R}(y_i)$ and $\mathcal{R}(z_i)$ are the distances used for the estimation of $\widehat{p_X}(x_i)$, $\widehat{p_Y}(y_i)$ and $\widehat{p_Z}(z_i)$ at the ith point, d_X, d_Y and d_Z are the dimensions of the signals X, Y and Z respectively. Similarly, using the maximum norm, we obtain

$$\mathcal{R}(x_i)=\mathcal{R}(z_i)\ \text{ and }\ \mathcal{R}(y_i)=\mathcal{R}(z_i). \tag{21}$$

Eq. (21) formally confirms (as suggested but not proved in [2]) that, if X and Y are independent, using the maximum norm and constraining the values $\mathcal{R}(x_i)$ and $\mathcal{R}(y_i)$ to be equal to $\mathcal{R}(z_i)$ allows to decrease the bias $\widehat{\mathcal{I}(X,Y)}-\mathcal{I}(X,Y)$. Eq. (20) extends this result when the Euclidean norm is used for the 3 individual spaces. We should mention that Eq. (18) no longer holds if signals X and Y are not independent. In this case only a part of the bias can be expected to be cancelled out.

So, finally, in the case of independence between X and Y, we introduced the following MI estimator

$$\widehat{\mathcal{I}(X,Y)} = -\frac{1}{n}\sum_{i=1}^{n}\left[\log\widehat{p_X}(x_i)+\log\widehat{p_Y}(y_i)-\log\widehat{p_Z}(z_i)\right] \tag{22}$$

with an (approximately) zero bias by choosing $\mathcal{R}(z_i)$ and by properly defining $\mathcal{R}(x_i)$ and $\mathcal{R}(y_i)$ using Eq. (20) or (21). When $\mathcal{R}(z_i)$ results from the kNN approach [i.e., when $\mathcal{R}(z_i)=\|k\mathrm{NN}\,(z_i)-z_i\|$ is the distance from z_i to its kth NN, also denoted $\mathcal{R}_k(z_i)$], this estimator is denoted by $\widehat{\mathcal{I}(X,Y)}^{k}_{\text{basic}}$ with Eq. (23) [with $k(z_i)=k$]. Hereafter, this estimator is written as $\widehat{\mathcal{I}(X,Y)}_{\text{basic,E}}$ for the Euclidean norm and by $\widehat{\mathcal{I}(X,Y)}_{\text{basic,M}}$ for the maximum norm, and called "basic estimator".

C. Bias reduction of MI estimation based on the new bias expression (X and Y dependent)

Now, to further eliminate the bias in MI estimation in the general case (X and Y are dependent), we consider again the estimation of individual entropies. Removing the bias $\mathcal{B}_X$ in Eq. (14) is not an easy task since its mathematical expression depends on the unknown probability density. However, we can expect to cancel it out considering a weighted linear combination [6]. Consequently, we introduce the following form of an ensemble estimator of entropy:

$$\widehat{\mathcal{H}(X)} = \left\{ -\frac{1}{n}\sum_{i=1}^{n}\left[(1-\alpha_i)\log\widehat{p_X}^{(1)}(x_i)\right]\right\} + \left\{ -\frac{1}{n}\sum_{i=1}^{n}\left[\alpha_i \log\widehat{p_X}^{(2)}(x_i)\right]\right\}, \tag{24}$$

where $\alpha_i, i = 1,..,n$ is a sequence of weighting coefficients to be determined, $\widehat{p_X}^{(1)}(\cdot)$ and $\widehat{p_X}^{(2)}(\cdot)$ are two density estimations $\left[\widehat{\frac{\int_{\mathcal{L}(x)} p_X(y)\mathrm{d}y}{v(x)}}\right]$ obtained from two distinct definitions of $\mathcal{L}(\cdot)$. Until now, $\mathcal{L}(x)$ was built either from a kNN approach or a KDE approach. In the first case, $\mathcal{R}(x)$ is deduced from the kth NN, and in the second case, $\mathcal{R}(x)$ depends on the imposed bandwidth. Hereafter, to carry on with the conjecture proposed in [2], we only consider the kNN approach integrating two steps (i) the choice of two different numbers of neighbors k_1 and k_2, (ii) the definition of the probability density estimators,

$$\widehat{p_X}^{(1)}(x_i) = \widehat{p_{k_1}}(x_i) \quad \text{and} \quad \widehat{p_X}^{(2)}(x_i) = \widehat{p_{k_2}}(x_i), \tag{25}$$

where

$$\widehat{p_{k_j}}(x) = \frac{k_j}{n\cdot v_{k_j}(x)}, \quad j = 1,2 \tag{26}$$

is the standard kNN density estimator as defined in [7]. The volume $v_k(x)$ is equal to the Lebesgue measure of $\mathcal{L}_k(x) = \{y : \|y - x\| \le \mathcal{R}_k(x)\}$, and $\mathcal{R}_k(x_i)$ is the distance between x_i and its kth NN.

Considering each bias term, we write

$$\mathcal{B}_X(x_i) \triangleq (1-\alpha_i)\mathcal{B}_{k_1}(x_i) + \alpha_i \mathcal{B}_{k_2}(x_i). \tag{27}$$

The question arises of how to choose α_i in Eq. (27) so that $\mathcal{B}_X(x_i) = 0$.

Given the Euclidean norm [Eq. (15)], we have

$$\begin{aligned}\mathcal{B}_X(x_i) &= (1-\alpha_i)\mathcal{B}_{k_1}(x_i) + \alpha_i\mathcal{B}_{k_2}(x_i) \\ &= \frac{(1-\alpha_i)\mathcal{R}_{k_1}^2(x_i) + \alpha_i\mathcal{R}_{k_2}^2(x_i)}{2\,(d_X+2)\,p(x_i)}\mathrm{tr}\left[\frac{\partial^2 p_X(x_i)}{\partial x^2}\right].\end{aligned} \tag{28}$$

Now, solving Eq. (28) for any $i = 1,..,n$ with respect to α_i leads to

$$\alpha_i = \frac{\mathcal{R}_{k_1}^2(x_i)}{\mathcal{R}_{k_1}^2(x_i) - \mathcal{R}_{k_2}^2(x_i)}. \tag{29}$$

When starting from Eq. (16) instead of Eq. (15) to address the maximum norm, Eq. (29) still holds. Practically, an optimal choice of the parameters k_1 and k_2 is not obvious. Nevertheless, it is possible to tune these two parameters to improve the original biased estimator.

In the dependent case we can apply the same strategy to X, Y and Z separately with distinct coefficients α_i^x, α_i^y, α_i^z and then compute the ensemble MI estimator using

$$\widehat{\mathcal{I}(X,Y)}_{\text{ens}} = \widehat{\mathcal{H}(X)}_{\text{ens}}^{k_1^x,k_2^x} + \widehat{\mathcal{H}(Y)}_{\text{ens}}^{k_1^y,k_2^y} - \widehat{\mathcal{H}(Z)}_{\text{ens}}^{k_1^z,k_2^z}, \tag{30}$$

where

$$\widehat{\mathcal{H}(U)}_{\text{ens}}^{k_1^u,k_2^u} = -\frac{1}{n}\sum_{i=1}^{n}\left[(1-\alpha_i^u)\log\frac{k_1^u}{n\cdot v_{k_1}(u_i)} + \alpha_i^u\log\frac{k_2^u}{n\cdot v_{k_2}(u_i)}\right], \tag{31}$$

with the pairs (k_1^x, k_2^x), (k_1^y, k_2^y) and (k_1^z, k_2^z) chosen independently for X, Y and Z.

In the independent case, the basic strategy [Eq. (22)] can be used. But we note that the values $\alpha_i^u = \frac{\mathcal{R}_{k_1}^2(u_i)}{\mathcal{R}_{k_1}^2(u_i)-\mathcal{R}_{k_2}^2(u_i)}$, with u replaced by x, y or z, are identical if we choose $\mathcal{R}_{k_1}^2$, $\mathcal{R}_{k_2}^2$ with the constraint imposed by Eq. (20) [or Eq. (21)].

Developing Eq. (30) with the substitution $\alpha_i^x = \alpha_i^y = \alpha_i^z = \alpha_i$, we get a mixed mutual information estimator

$$\widehat{\mathcal{I}(X,Y)}_{\text{mixed}}^{k_1,k_2} = \widehat{\mathcal{H}(X)}_{\text{mixed}}^{k_1^x,k_2^x} + \widehat{\mathcal{H}(Y)}_{\text{mixed}}^{k_1^y,k_2^y} - \widehat{\mathcal{H}(Z)}_{\text{mixed}}^{k_1^z,k_2^z}, \tag{32}$$

where

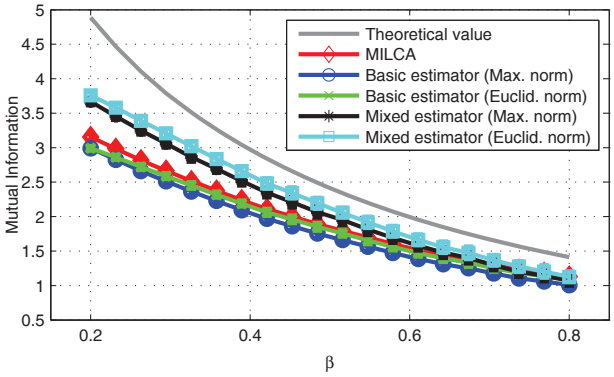
$$\widehat{\mathcal{H}(U)}_{\text{mixed}}^{k_1^u,k_2^u} = -\frac{1}{n}\sum_{i=1}^{n}\left[(1-\alpha_i)\log\frac{k_{k_1}(u_i)}{n\cdot v_{k_1}(u_i)} + \alpha_i\log\frac{k_{k_2}(u_i)}{n\cdot v_{k_2}(u_i)}\right]. \tag{33}$$

In summary, our mixed MI estimator is built following the three steps:

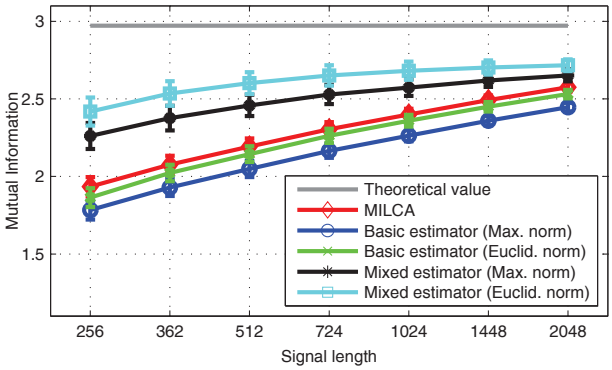
(i) Fix the number of NNs (k_1 and k_2 separately) in the joint space $\mathcal{S}_Z$ to get the distances between the center point z_i and the particular NNs (k_1th NN and k_2th NN), marked as $\mathcal{R}_{k_1}(z_i)$ and $\mathcal{R}_{k_2}(z_i)$

(ii) Use $\mathcal{R}_{k_1}(z_i)$ and $\mathcal{R}_{k_2}(z_i)$ to get respectively $\mathcal{R}_{k_1}(x_i)$, $\mathcal{R}_{k_1}(y_i)$, and $\mathcal{R}_{k_2}(x_i)$, $\mathcal{R}_{k_2}(y_i)$, using Eq. (20) or Eq. (21) (depending on the norm) and determine the numbers of points $k_{k_1}(x_i)$, $k_{k_1}(y_i)$, $k_{k_2}(x_i)$ and $k_{k_2}(y_i)$ falling into the corresponding regions

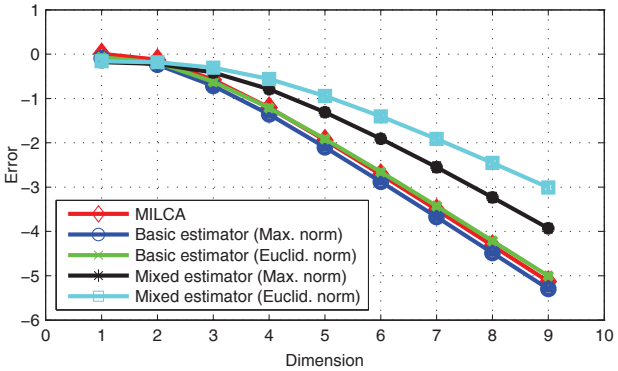
(iii) Estimate $\mathcal{H}(X)$ and $\mathcal{H}(Y)$ with Eq. (33), where α_i is given by Eq. (29), $\mathcal{H}(Z)$ being calculated similarly [with $k_{k_1}(z_i) = k_1$ and $k_{k_2}(z_i) = k_2$] and then calculate $\widehat{\mathcal{I}(X,Y)}_{\text{mixed}}^{k_1,k_2}$ by Eq. (32). The resulting estimator is named "mixed estimator" and



(a) (Color Online) Mutual information (in nats) estimated with varying β, $d = 3$, $n = 512$.



(b) (Color Online) Mutual information (in nats) estimated with different signals lengths, $\beta = 0.4$, $d = 3$.



(c) (Color Online) Mean estimation error $\overline{\widehat{\mathcal{I}(X,Y)} - \mathcal{I}(X,Y)}$ (in nats) with varying dimension, $\beta = 0.5$, $n = 512$.

FIG. 1. Mutual information and mean estimation error using the different estimators $\widehat{\mathcal{I}(X,Y)}_{\text{basic,E}}$, $\widehat{\mathcal{I}(X,Y)}_{\text{basic,M}}$, $\widehat{\mathcal{I}(X,Y)}_{\text{mixed,E}}$ and $\widehat{\mathcal{I}(X,Y)}_{\text{mixed,M}}$ with 100 trials.

denoted by $\widehat{\mathcal{I}(X,Y)}_{\text{mixed,E}}$ for the Euclidean norm and $\widehat{\mathcal{I}(X,Y)}_{\text{mixed,M}}$ for the maximum norm.

Note that $\widehat{\mathcal{I}(X,Y)}^{k}_{\text{basic}}$ is obtained by replacing Eq. (32) by Eq. (23) in step (iii).

III. NUMERICAL TEST

The following linear model is generated

$$Y = X + \beta \cdot e, \tag{34}$$

where X and e are independent d-dimensional random vectors, and both of them follow a zero mean Gaussian distribution $N(0, \mathrm{I})$ (I is the identity matrix). Clearly, when β decreases the dependence between X and Y increases. The theoretical value of the mutual information $\mathcal{I}(X,Y)$ is equal to $\frac{d}{2}\log\left(\frac{1+\beta^2}{\beta^2}\right)$. For simulations we use sequences of n independent samples $(X_i, Y_i)\,, i = 1, .., n$ from the distribution of (X,Y).

We test the 4 estimators $\widehat{\mathcal{I}(X,Y)}_{\text{basic,E}}$, $\widehat{\mathcal{I}(X,Y)}_{\text{basic,M}}$, $\widehat{\mathcal{I}(X,Y)}_{\text{mixed,E}}$ and $\widehat{\mathcal{I}(X,Y)}_{\text{mixed,M}}$ to estimate $\mathcal{I}(X,Y)$. We also run the MI estimator algorithm freely available from the MILCA toolbox [8], simply denoted by MILCA, and which a priori corresponds to $\widehat{\mathcal{I}(X,Y)}_{\text{basic,M}}$ [9].

Throughout the experimentation, we choose $k = 6$ (the default k value of MILCA toolbox) for the basic estimators, and $k_1 = 6$, $k_2 = 20$ for the mixed estimators. The statistical mean and variance of the five estimators are estimated by an averaging on 100 trials.

Fig. 1(a) displays the performance of the five algorithms, for a given dimension ($d = 3$), a number of points equal to $n = 512$, and different values of β. The correlation between X and Y is all the more important as β is low. It comes out that all estimators are comparable when β reaches 0.8 (corresponding to a correlation coefficient around 0.78 between same ranks coordinates

of X and Y). When the signals are highly correlated (low values of β), the basic estimators still show identical behaviors, but, in this case, the two new mixed estimators clearly outperform the former whatever the norm, the best result being obtained using the Euclidean norm based estimator. Even if all results are not presented here, we find that the two new estimators outperform the basic ones using either $k = 6$ or $k = 20$.

We also tested the five estimators for different lengths of the time series for given values of β and d. As displayed in Fig. 1(b), the two new mixed estimators behave better whatever the length of the signals (ranging from 512 until 2048), the improvement being all the more important that the signal length is short.

When computing the error between the different estimators and the theoretical value, for a given value of β ($\beta = 0.5$) corresponding to a correlation coefficient between the signals equal to 0.89, and an increasing dimension, the same conclusion globally holds, as displayed in Fig. 1(c). The new mixed estimators clearly outperform the basic ones (which display comparable behavior) especially for high dimensions. However, for very low dimensions ($d = 1$ or $d = 2$), the original estimators may be preferred. Clearly, for all estimators, the error grows along with the dimension, the best result being systematically obtained with the mixed estimator based on the Euclidean norm. Since the standard deviations are quite low, they are not shown in these figures. Using the basic estimators (or MILCA), the standard deviation varies from 0.03 to 0.06 which is extremely low compared to the estimated values of mutual information (approximately from 1 to 5). As for the mixed estimators, the standard deviation varies from 0.04 to 0.09. The increasing in standard deviation can be considered as negligible in comparison to the accuracy of the estimation.

IV. SUMMARY

In this paper, we investigated the difficult issue of bias reduction on mutual information estimation. Once we established a relation between the systematic bias and the distance parameter for plug-in entropy estimator, two strategies, a basic one and a new one involving mixed estimators, were discussed. Experimental results allowed us to assess the performance of the new estimators using Euclidean or maximum norms to get a more accurate estimation of mutual information.

[1] T. M. Cover and J. A. Thomas, *Elements of information theory* (John Wiley & Sons, New York, 1991).

[2] A. Kraskov, H. Stögbauer, and P. Grassberger, Phys. Rev. E **69**, 066138 (2004).

[3] S. Frenzel and B. Pompe, Phys. Rev. Lett. **99**, 204101(2007).

[4] Q. Wang, S. R. Kulkarni, and S. Verdú, IEEE Trans. Inf. Theory **55**, 2392 (2009).

[5] K. Fukunaga and L. Hostetler, IEEE Trans. Inf. Theory **19**, 320 (1973).

[6] K. Sricharan, D. Wei, and A. O. Hero III, IEEE Trans. Inf. Theory **59**, 4374 (2013).

[7] K. Fukunaga, *Introduction to statistical pattern recognition*, 2nd edition (Academic Press, San Diego, 1990).

[8] http://www.ucl.ac.uk/ion/departments/sobell/Research/RLemon/MILCA/MILCA

[9] The algorithm encoded in the toolbox [Eq. (2) or (3)] is not documented.